haberman dataset EDA plotting

In [1]:
```python
import warnings
warnings.filterwarnings("ignore")
```

In [20]:
```python
#haberman dataset EDA plotting
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
#Load haberman.csv into a pandas dataFrame.
hman = pd.read_csv("haberman.csv")
```

In [21]:
```python
#tells the shape of given csv file
print(hman.shape)
```

```
(306, 4)
```

In [22]:
```python
#columns present in file
print(hman.columns)
```

```
Index(['Age', 'Op_Year', 'axil_nodes_det', 'Surv_status'], dtype='object')
```

In [5]:
```python
#no. of patients alive after 5 year of cancer treatment and patients who cant survive
hman["Surv_status"].value_counts()
```

Out[5]:
```
1    225
2     81
Name: Surv_status, dtype: int64
```

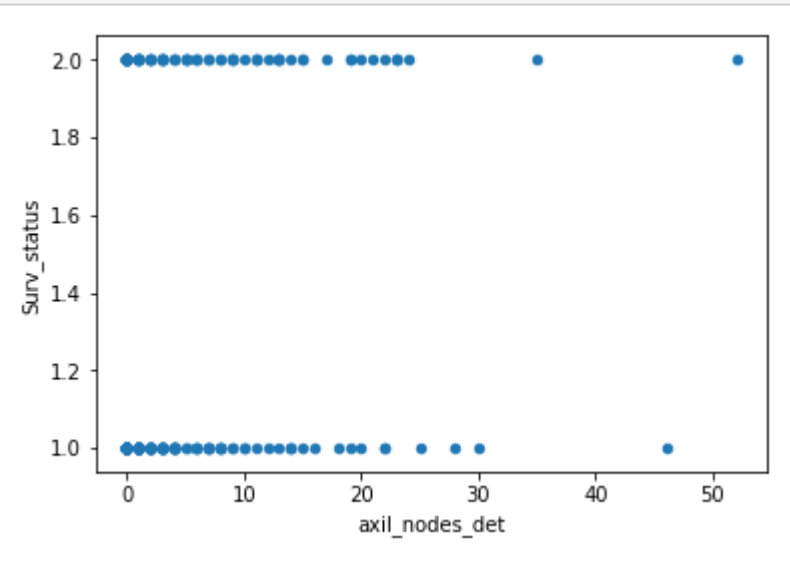
In [24]:
```python
hman["Surv_status"][hman["Surv_status"]==1]='Yes'
hman["Surv_status"][hman["Surv_status"]==2]='No'
```

observation

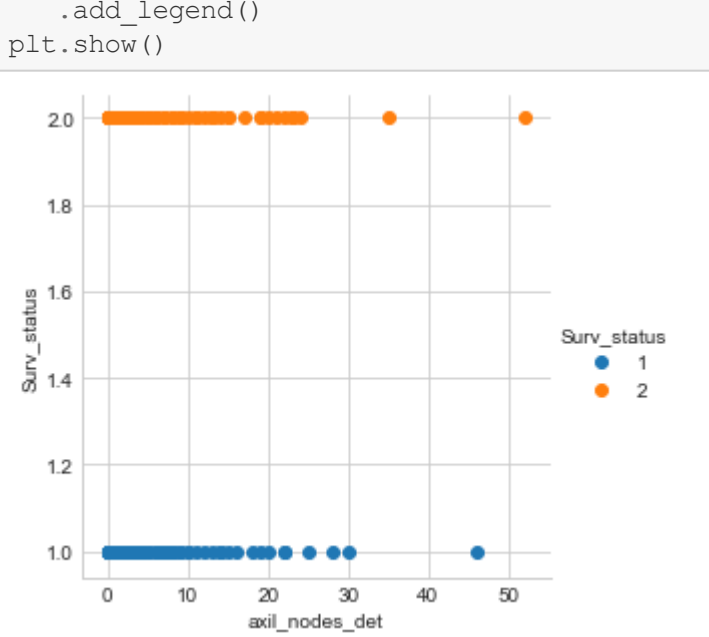
1. Here "surv_counts" shows the survival of patients after 5 year of cancer treatment.
2. "1" shows that patient survive 5 year or more .
3. "2" shows that patient can not survive 5 year.
4. according to this data there are 225 people survives and 81 can not survive.

2-D scatter plot

In [11]:
```python
#2-D scatter plot of survey status  of patients w.r.t. axillary nodes detected
hman.plot(kind="scatter",x="axil_nodes_det",y="Surv_status")
plt.show()
```



In [12]:
```python
#2-D scatter plot of survey status  of patients w.r.t. axillary nodes detected in different hue
import seaborn as sns
sns.set_style("whitegrid")
sns.FacetGrid(hman,hue="Surv_status",size=4)\
   .map(plt.scatter,"axil_nodes_det","Surv_status")\
   .add_legend()
plt.show()
```



Pair plot

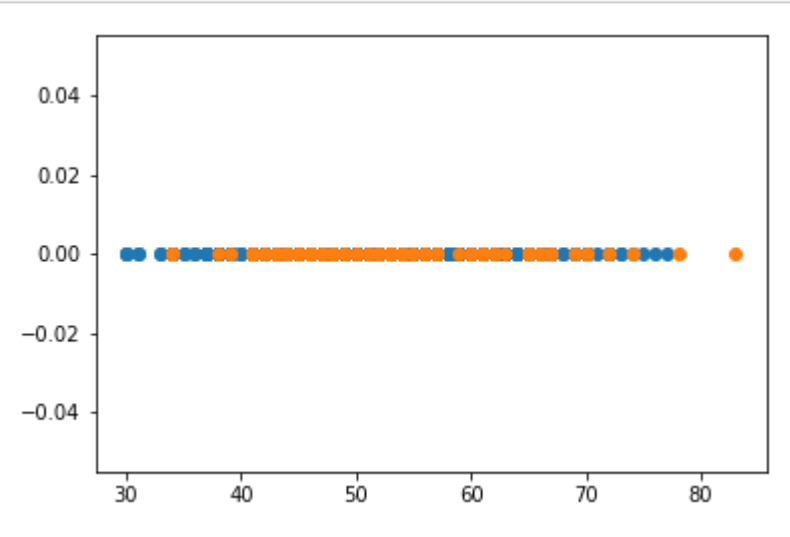
In [26]:
```python
#pair plot of the give data of hyberman
plt.close()
sns.set_style("whitegrid")
sns.pairplot(hman,hue="Surv_status",size=3)
plt.show()
```

Observation

1.This is the all required pair plot which show survival of peitients. 2.as we can see in year section that in every year during analysis there are both some peaple survived 5 year and some peaple cant. 3.there are some more assumptions we can make after analysing these plottings

1-D plotting

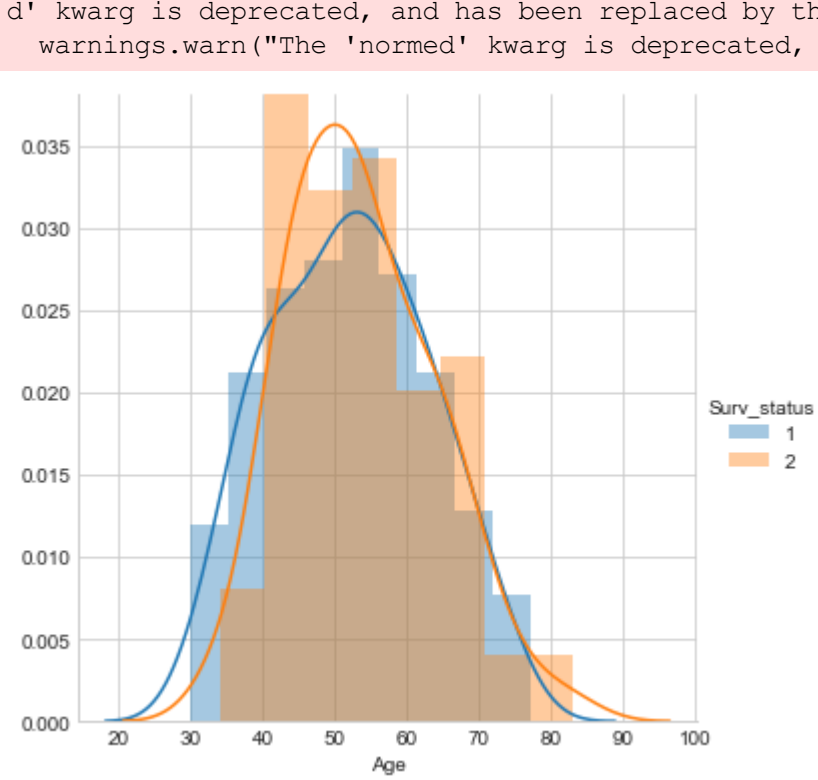
In [8]:
```python
#1-D scatter plot of Age of patients with respect to their survey status
import numpy as np
hman_surv = hman.loc[hman["Surv_status"] == 1]
hman_nsurv = hman.loc[hman["Surv_status"] == 2]

plt.plot(hman_surv["Age"],np.zeros_like(hman_surv['Age']),"o")
plt.plot(hman_nsurv["Age"],np.zeros_like(hman_nsurv['Age']),"o")

plt.show()
```



1-d ditributed plotting or histogram with PDF of survey status of patient w.r.t. their Ages

In [21]:
```python
#1-d ditributed plotting or histogram with PDF of survey status of patient w.r.t. their Ages
sns.FacetGrid(hman,hue="Surv_status",size=5) \
   .map(sns.distplot,"Age") \
   .add_legend()
plt.show()
```

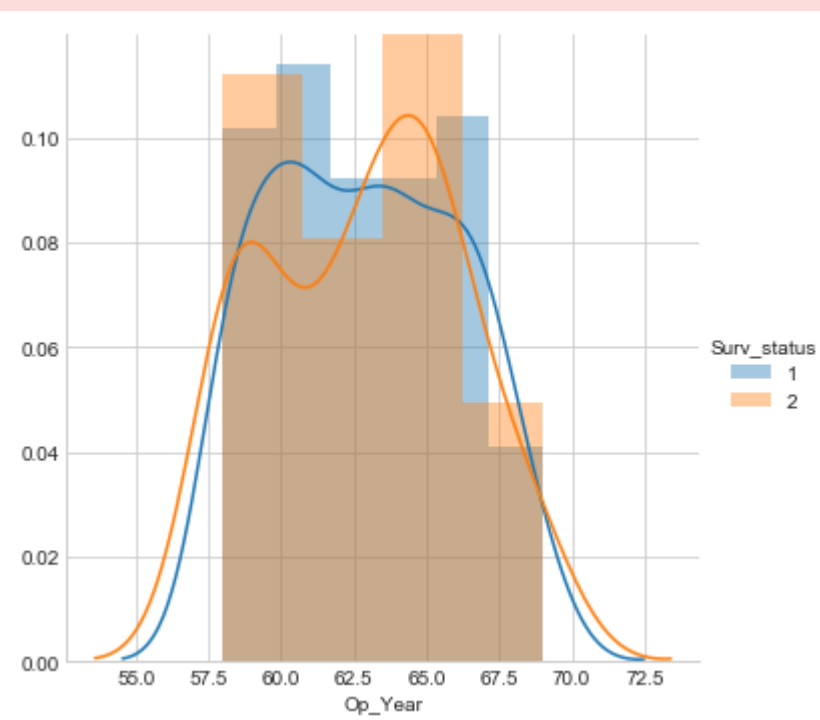
```
C:\Users\LiGht\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'norme
d' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\LiGht\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'norme
d' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```



1-d ditributed plotting or histogram with PDF of survey status of patient w.r.t. their operation year

In [25]:
```python
#1-d ditributed plotting or histogram with PDF of survey status of patient w.r.t. their operation ye
ar
sns.FacetGrid(hman,hue="Surv_status",size=5) \
   .map(sns.distplot,"Op_Year") \
   .add_legend()
plt.show()
```

```
C:\Users\LiGht\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'norme
d' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\LiGht\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'norme
d' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```



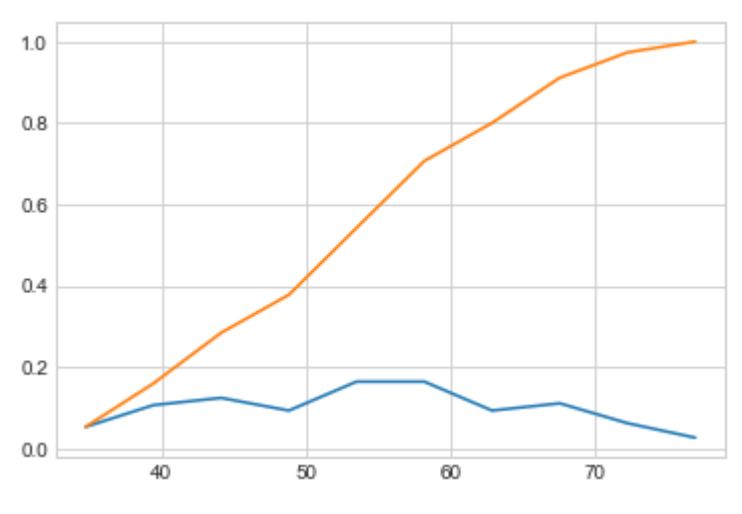
1-d ditributed plotting or histogram with PDF of survey status of patient w.r.t. their axillary nodes detected

In [ ]:
```python
# 1-d ditributed plotting or histogram with PDF of survey status of patient w.r.t. their axillary no
des detected
sns.FacetGrid(hman,hue="Surv_status",size=5) \
   .map(sns.distplot,"axil_nodes_det") \
   .add_legend()
plt.show()
```

pdf and cdf of patients who survived after 5 year of cancer treatment

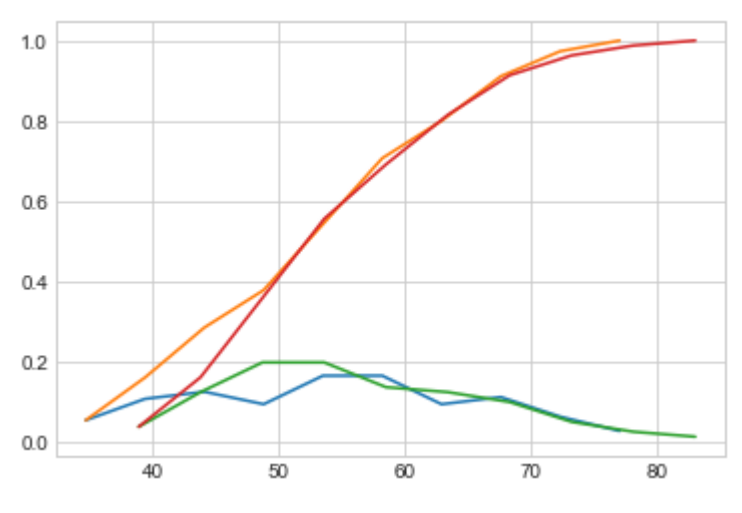
In [23]:
```python
# PDF and CDF of given data
#pdf and cdf of patients who survived after 5 year of cancer treatment
counts,bin_edges=np.histogram(hman_surv["Age"],bins=10,density=True)
pdf=counts/(sum(counts))
print(pdf)
print(bin_edges)
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)
plt.show()
```

```
[0.05333333 0.10666667 0.12444444 0.09333333 0.16444444 0.16444444
 0.09333333 0.11111111 0.06222222 0.02666667]
[30.  34.7 39.4 44.1 48.8 53.5 58.2 62.9 67.6 72.3 77. ]
```



In [25]:
```python
#pdf and cdf of patients who survived after 5 year of cancer treatment
counts,bin_edges=np.histogram(hman_surv["Age"],bins=10,density=True)
pdf=counts/(sum(counts))
print(pdf)
print(bin_edges)
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:], cdf)

##pdf and cdf of patients who cant survived after 5 year of cancer treatment
counts,bin_edges=np.histogram(hman_nsurv["Age"],bins=10,density=True)
pdf=counts/(sum(counts))
print(pdf)
print(bin_edges)
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:], cdf)

plt.show()
```

```
[0.05333333 0.10666667 0.12444444 0.09333333 0.16444444 0.16444444
 0.09333333 0.11111111 0.06222222 0.02666667]
[30.  34.7 39.4 44.1 48.8 53.5 58.2 62.9 67.6 72.3 77. ]
[0.03703704 0.12345679 0.19753086 0.19753086 0.13580247 0.12345679
 0.09876543 0.04938272 0.02469136 0.01234568]
[34.  38.9 43.8 48.7 53.6 58.5 63.4 68.3 73.2 78.1 83. ]
```



Observation

1.patients who's cancer treatment done within age of 33, they are survived after 5 year of treatment 2.patients who's cancer treatment done after age of 78, they can't survived after 5 year of treatment

Mean, variance and standard deviantion

In [33]:
```python
# mean of given data
print("Means:")
print(np.mean(hman_surv["Age"]))
print(np.mean(hman_nsurv["Age"]))
# standard deviation of given data
print("\nstandard deviation:")
print(np.std(hman_surv["Age"]))
print(np.std(hman_nsurv["Age"]))
```

```
Means:
52.01777777777778
53.67901234567901

standard deviation:
10.98765547510051
10.10418219303131
```

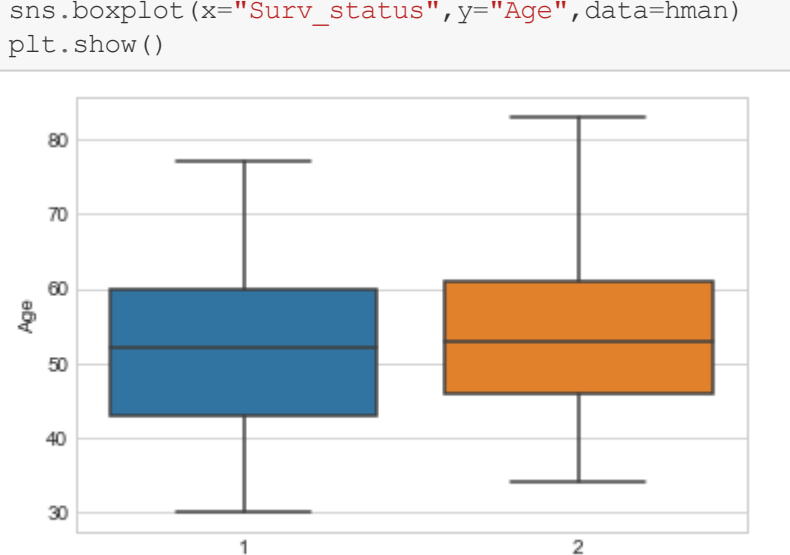
Median, Percentile, Quantile, IQR, MAD

In [41]:
```python
# median of given data
print("medians:")
print(np.median(hman_surv["Age"]))
print(np.median(hman_nsurv["Age"]))
# median of given data
print("\n 90th Percentile:")
print(np.percentile(hman_surv["Age"],90))
print(np.percentile(hman_nsurv["Age"],90))
# quantiles of given data
print("\nquantiles:")
print(np.percentile(hman_surv["Age"],np.arange(0,100,25)))
print(np.percentile(hman_nsurv["Age"],np.arange(0,100,25)))
# MAD of given data
from statsmodels import robust
print("\nMAD:")
print(robust.mad(hman_surv["Age"]))
print(robust.mad(hman_nsurv["Age"]))
```

```
medians:
52.0
53.0

 90th Percentile:
67.0
67.0

quantiles:
[30. 43. 52. 60.]
[34. 46. 53. 61.]

MAD:
13.343419966550417
11.860817748044816
```

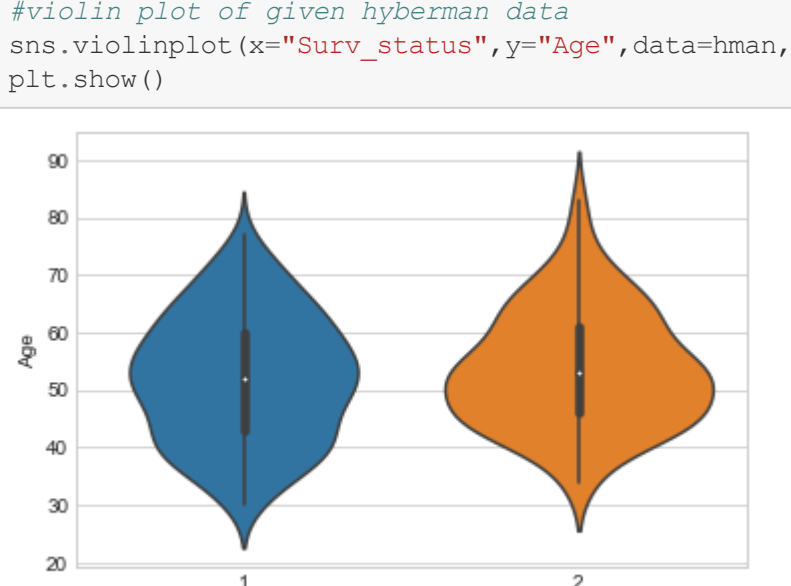
Box plot and Whiskers

In [45]:
```python
#box_plot and whiskers to show survey
sns.boxplot(x="Surv_status",y="Age",data=hman)
plt.show()
```



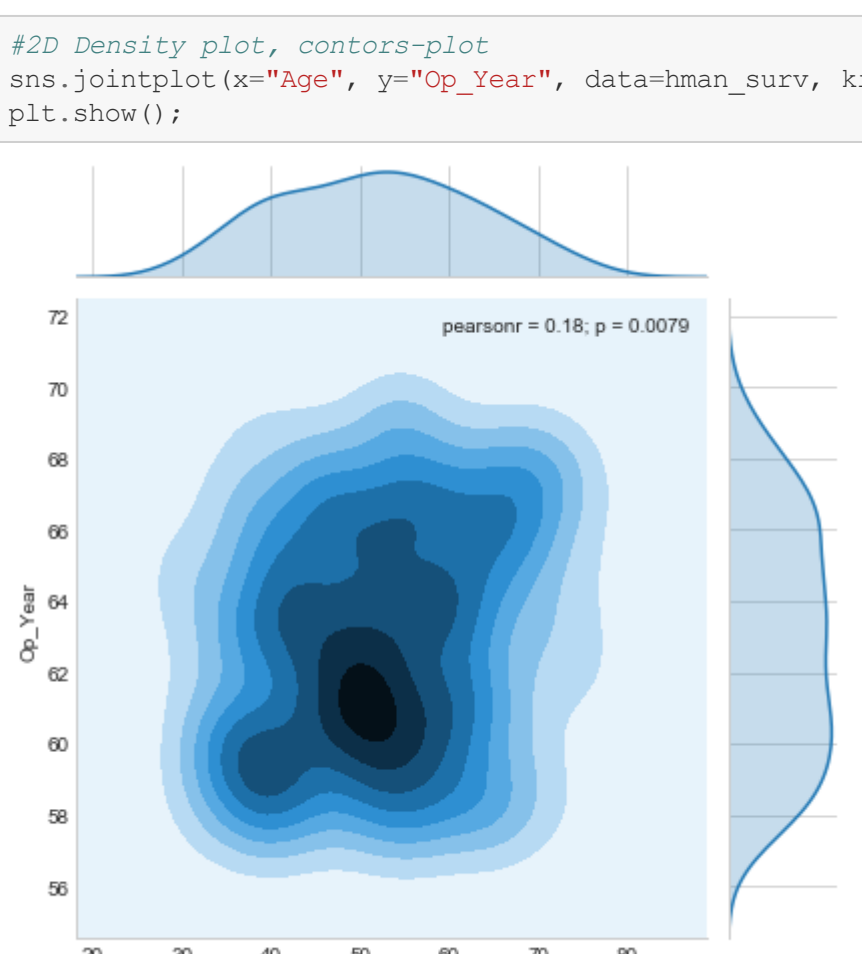
Violin plots

In [46]:
```python
#violin plot of given hyberman data
sns.violinplot(x="Surv_status",y="Age",data=hman,size=10)
plt.show()
```



2-D contour plot

In [47]:
```python
#2D Density plot, contors-plot
sns.jointplot(x="Age", y="Op_Year", data=hman_surv, kind="kde");
plt.show();
```



Final Observation